What if Othello-Playing Language Models Could See?

Xinyi Chen * [1] Yifei Yuan * [2] Jiaang Li [3] Serge Belongie [3] Maarten de Rijke [1] Anders Søgaard [3]

Abstract

Language models are often said to face a symbol grounding problem. While some argue that world understanding can emerge from text alone, others suggest grounded learning is more efficient. We explore this through Othello, where the board state defines a simplified, rule-based world. Building on prior work, we introduce VISOTHELLO, a multi-modal model trained on move histories and board images. Using next-move prediction, we compare it to mono-modal baselines and test robustness to semantically irrelevant perturbations. We find that multi-modal training improves both performance and the robustness of internal representations. These results suggest that grounding language in visual input helps models infer structured world representations.

1. Introduction

Does a language model truly understand what cat refers to? While no one fully grasps the essence of a cat, human language users know enough to use the word appropriately—recognizing cats in images, inferring them from descriptions, and using the term naturally in context. Whether mono-modal language models can achieve this level of grounding remains an open question.

Rather than debating whether symbol grounding is *in principle* impossible for mono-modal models (Mitchell & Krakauer, 2023; Mollo & Millière, 2023), this paper focuses on a more pragmatic hypothesis: that incorporating multiple modalities enables more efficient learning. This question is orthogonal to, but compatible with, the idea that mono-modal models can still induce a form of referential semantics (Søgaard, 2023; Huh et al., 2024).

To explore this hypothesis, we turn to the concept of world model—a system that learns to represent the latent state of an environment and predict its evolution over time. While many world model studies use high-dimensional or continuous environments (e.g., robotics or simulated physics), we use the game of Othello as a minimal, symbolic world with deterministic transitions and well-defined rules.

Prior work has used this setup to investigate emergent world representations by training and probing models—from small transformers (Li et al., 2023; Nanda et al., 2023) to large language models (LLMs) (Yuan & Søgaard, 2025)—on sequences of Othello moves. These studies provide evidence that language models can learn to track the board state, forming a rudimentary world model when trained on large-scale sequential data, though this learning remains confined to the text-only domain.

We extend the Othello learning task to a multi-modal setting by introducing **VISOTHELLO**, a multi-modal Othello model trained on sequences of move histories and their corresponding board images; see Figure 1. For each sequence of moves, we generate a corresponding sequence of board state images, with each image depicting the board at a specific time step. We then apply masking strategies to selected move tokens and train the model to predict the missing steps, using both the move history and associated visual context.

Our main goal is to investigate whether access to visual state information enhances sample efficiency and accelerates learning. We break down the main research questions into several related aspects:

Main question	Is multi-modal (Othello) learning faster?
Sub$_1$	Is multi-modal learning better?
Sub$_2$	Is multi-modal grounding better?

To address **Sub**$_1$, we compare VISOTHELLO against several baselines on the task of *next move prediction*, where the model predicts the next token given a partial game sequence. We evaluate the performance across varying data scales to assess learning efficiency. For **Sub**$_2$, we perform a semantically irrelevant perturbation analysis by rotating the board image during inference, assessing whether the models trained on original images remain robust and continue to predict legal moves accurately.

Our findings show that multi-modal training leads to better performance and faster learning than learning from text

*Equal contribution [1]University of Amsterdam [2]ETH Zürich [3]University of Copenhagen. Correspondence to: Xinyi Chen <x.chen2@uva.nl>, Yifei Yuan <yuanyif@ethz.ch>.

ICML 2025 Workshop on Assessing World Models.

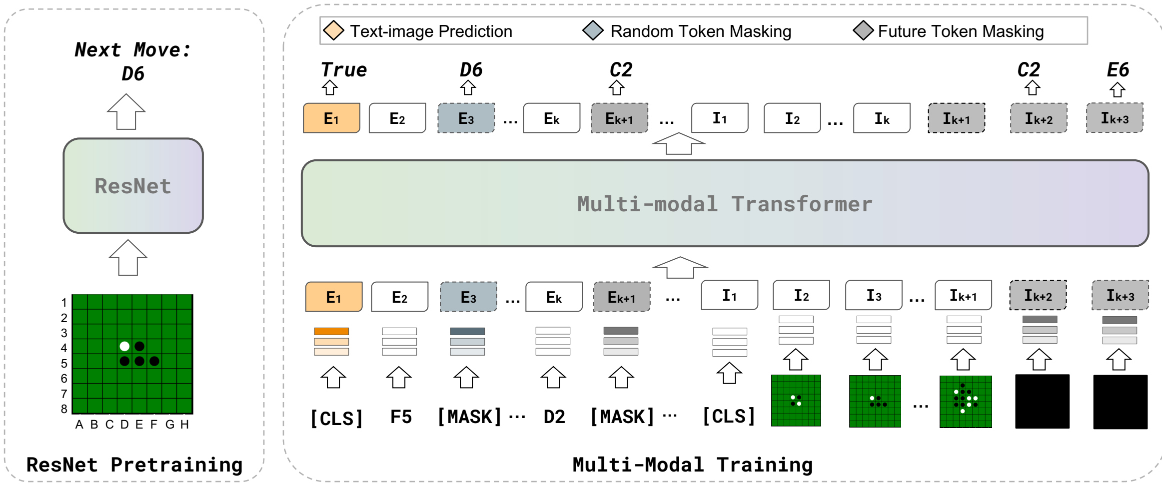


Figure 1. Architecture of VISOTHELLO. The model integrates visual and textual inputs by encoding board images and corresponding move sequences using a Transformer. During pretraining, (i) a ResNet is trained to predict the next move from the current board image; (ii) the multi-modal Transformer is pretrained with three objectives: text-image prediction, random token masking, and future token masking.

alone. In addition, multi-modal models exhibit greater robustness under semantically irrelevant perturbations, as demonstrated in our board rotation analysis. These results suggest that while language models can infer world state representations from text alone, grounding them with visual input enables more efficient and robust construction of such representations.

2. Multi-modal Othello Training

2.1. Training Paradigm

Different from prior works that train Othello models in an autoregressive manner by predicting moves step-by-step, we adopt a **BERT-style masked language modeling** approach for training VISOTHELLO. This avoids the computational overhead and complexity of autoregressive generation, enabling efficient bidirectional reasoning over static visual-text inputs without framing the task as video modeling (for a detailed explanation, see Appendix E). Specifically, we train VISOTHELLO based on VisualBERT (Li et al., 2019).

2.2. Input Representation

Textual input. Following prior works (Li et al., 2023; Karvonen et al., 2024), we represent each game as a sequence of moves, where each move at time step t is treated as a token, denoted as m_t. Our vocabulary consists of 64 unique tokens, corresponding to the 64 tiles on the board; see Fig 1.
Image input. In addition to the textual input, we provide the model with a sequence of corresponding board images. As demonstrated in Figure 1, each image b_t represents the board state after moves $m_1, m_2, \ldots, m_{t-1}$, and serves as visual context for predicting the next move m_t. To extract visual features, considering the differences between Othello board images and object images in ImageNet (Russakovsky et al., 2015), we pretrain an Othello-specific image encoder using a ResNet-18 backbone (He et al., 2016) to extract visual features:

$$\mathbf{v}_t = \phi(\mathbf{b}_t) \in \mathbb{R}^{d_v},$$

where ϕ denotes the image encoder, and d_v is the dimensionality of the visual representation. The image embeddings v_t are treated as image tokens to the input of models and are separated from the text tokens by a special token *[SEP]*.

2.3. VISOTHELLO Training

We train the VISOTHELLO model using two types of masked language modeling (MLM) strategies to enhance its ability to learn meaningful representations of both textual and visual game sequences. MLM enables the model to develop a deeper understanding of game dynamics. The overall training objective $\mathcal{L}_{\text{total}}$ is defined as the sum of the masked modeling and the text-image prediction loss as follows.

Random token masking. Following the training setup of BERT and VisualBERT, we apply random masking to the move sequence with an 80% probability, masking 15% of the move tokens at random, while keeping the image tokens fully visible. With the random masking task, the model learns to infer missing information using both modalities, reinforcing cross-modal alignment.

Train Size	0	1k	3k	5k	10k	20k
Othello-GPT	7.34	19.61	70.89	80.00	82.56	81.46
BERT-S	20.74	90.86	90.81	90.79	91.78	93.09
ResNET-18-S	6.71	60.78	71.27	75.58	81.74	88.90
VISOTHELLO-S	23.64	90.82	**93.87**	**93.25**	**94.23**	93.57
BERT-P	20.74	90.35	91.33	91.91	92.14	92.65
ResNET-18-P	6.71	71.28	84.57	88.08	91.02	92.23
VISOTHELLO-P	**43.30**	**91.42**	92.15	92.89	93.39	**94.03**

Table 1. Legal move accuracy (%) for next move prediction across different data sizes. We highlight the best performing model of each training set size in bold. $-P$ indicates the model is pretrained, while $-S$ indicates it is trained from scratch.

Figure 2. Illustration of probing results for BERT and VISOTHELLO trained with different dataset set sizes.

Future token masking. To align with the next-move prediction setup used in Othello-GPT (Li et al., 2023), we additionally apply future token masking to the game sequence with a 20% probability. Given a textual move sequence of $m_1, m_2, \ldots, m_s$, we randomly select a step t ($1 \leq t \leq s$) as the prediction target, and then mask all future tokens from m_t, which are $m_t, m_{t+1}, \ldots, m_s$. To prevent information leakage, we also mask all the image tokens that contains the future move information, which is $v_{t+1}, v_{t+2}, \ldots, v_s$. This setup reduces dependence on bidirectional context and fits better with the next-move prediction setup in the Othello.

Text-image prediction. We also adapt the sentence-image prediction task in original VisualBERT training for the Othello task. For a given sequence of image tokens, we replace the corresponding move sequence to a random sequence at a chance of 50%. The model is trained to distinguish whether the text and image sequences are from the same game via binary classification. This helps to train the model better learn implicit alignments between language and vision.

3. Experiments

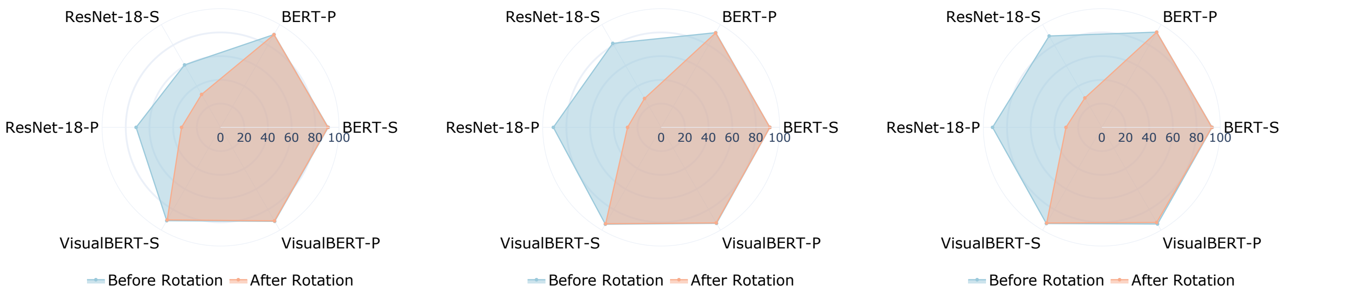
3.1. Experimental Setups

To better assess the impact of multi-modal learning, we include text-only baselines such as Othello-GPT, BERT and vision-only model ResNet for direct comparison. We collect a total of 25,657 real game records from the EOTHELLO website[1], which serve as the textual sequence inputs for our dataset. We generate the corresponding board images based on the dataset and split the dataset into training (80%), validation (5%), and test (15%) sets. When evaluating VISOTHELLO, we use **legal move accuracy**. It measures whether the predicted move m_t, given the move history m_1, $m_2, \ldots, m_{t-1}$, is valid under Othello's rules. To assess the learning efficiency of mono-modal and multi-modal models, we train all models on the full dataset (20k samples) as well as on randomly sampled subsets of 1k, 3k, 5k, and 10k examples. For VISOTHELLO, we use the best performing image encoder, ResNet-18 pretrained and fine-tuned on the full 20k dataset, for feature extraction. We also investigate the impact of pretraining by training each model either from scratch or from publicly available pretrained weights. Details see Appendix F and G.

3.2. Experimental Results

Table 1 report the next legal move prediction accuracy of various models across different dataset sizes. Several key observations emerge from these results. **Multi-modal learning is more sample efficient.** VISOTHELLO achieves high accuracy (over 91% in next legal move prediction) with as few as 1k training examples, while uni-modal models either require more data or fail to reach the same performance ceiling. This suggests that multi-modal learning is more sample-efficient—VISOTHELLO shows stronger performance at smaller scales than uni-modal baselines. This observation aligns with previous work from Zhuang et al. (2024). **Pretraining information is not consistently helpful.** While pretraining improves performance in some cases, especially in low-data regimes, its effect is not consistent across modalities. For ResNet-18, pretraining provides a substantial boost at smaller dataset sizes, but the improvements diminish as more data becomes available. In contrast, for both BERT and VISOTHELLO, pretraining does not consistently lead to significant gains across training sizes.

[1] https://www.eothello.com/



(a) Legal move acc. with 1K samples (b) Legal move acc. with 10K samples (c) Legal move acc. with 20K samples

Figure 3. Comparison of models' performance with and without board rotation across different training dataset sizes. The results demonstrate that multi-modal models maintain better performance under rotation compared to purely visual models. $-P$ indicates the model is pretrained, while $-S$ indicates it is trained from scratch.

This observation is consistent with prior findings by Yuan & Søgaard (2025), suggesting linguistic pretraining may offer limited benefit for rule-based environments such as Othello.

3.3. Probing Internal Representations

To assess whether VISOTHELLO learns meaningful internal representations of the board state, we train a linear probe to predict the state of each tile—i.e., whether it is empty, contains the player's disc, or the opponent's disc—based on hidden activations after processing a move sequence, following the approach of Nanda et al. (2023).

Results. Figure 2 shows F1 scores from linear probes trained to predict tile-level board states from selected layers of BERT and VISOTHELLO. When models are randomly initialized (0 examples), VISOTHELLO already encodes more board-relevant structure than BERT, achieving substantially higher probe performance in early layers. This may be attributed to the use of a ResNet encoder pretrained on Othello board images, which already encodes useful spatial structure. As training dataset increases, both models improve, but VISOTHELLO consistently achieves higher scores—especially in deeper layers. After training on 20k examples, VISOTHELLO reaches 77.55 F1 at Layer 18, compared to 62.28 for BERT. This suggests that VISOTHELLO learns more accurate internal representations of the board state, benefiting from both multi-modal input and architectural modifications.

4. Semantically Irrelevant Perturbation

We focus on board rotation as a concrete instance, which each test board is rotated 180 degrees. As illustrated in Figure 4, this transformation corresponds to a spatial inversion: for image-based models, this involves rotating the game board in the input image; for language-based models, it requires flipping the row and column indices of the move representation (e.g., D3 becomes E6, resulting in a different move token ID). We apply the rotation **only at test time**, evaluating models that were trained on the original (unrotated) training data. All models are assessed on their ability to predict the next legal move, as described in Section 3. This allows us to examine whether models rely on absolute visual or positional cues, or whether they have learned more abstract, generalizable representations of the board state.

As shown in Figure 3, BERT remains robust under board rotation, maintaining 90–93% accuracy across settings. This is expected, as symbolic move sequences can be deterministically remapped (e.g., D3 to E6). In contrast, ResNet-18's accuracy drops sharply to 28–35%, indicating a failure to learn rotation-invariant representations and a reliance on absolute spatial patterns. Without access to move history or turn information, it depends on ambiguous visual cues easily disrupted by rotation—a key limitation of purely visual models in tasks like Othello. VISOTHELLO, combining ResNet's visual features with BERT's move encoding, sustains high accuracy (91–93%) even after rotation. Sequence information, including player turns and prior moves, helps disambiguate visual input, while language guidance stabilizes predictions under spatial transformations. This highlights the advantage of multi-modal grounding: *by aligning perceptual input with symbolic context,* VISOTHELLO *overcomes the spatial brittleness of vision-only models.*

We then perform feature alignment in Appendix J.

5. Conclusion

We address the task of learning to play Othello and extend it to a multi-modal setting by introducing a novel model, VISOTHELLO. Specifically, we examine whether access to visual state information improves sample efficiency and accelerates learning by comparing VISOTHELLO to existing mono-modal baselines on next move prediction. To further assess the benefits of multi-modal grounding, we introduce a board rotation perturbation and conduct feature alignment analysis to evaluate whether the models learn more robust and aligned representations. Our findings suggest that grounding language models with visual input leads to more efficient and stable learning, validating the world model theory (detailed discussion see Appendix A and B).

References

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Caucheteux, C., Gramfort, A., and King, J.-R. Deep language algorithms predict semantic comprehension from brain activity. *Scientific reports*, 12(1):16327, 2022.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. Word translation without parallel data. *The Sixth International Conference on Learning Representations*, 2018.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Feng, X., Luo, Y., Wang, Z., Tang, H., Yang, M., Shao, K., Mguni, D., Du, Y., and Wang, J. Chessgpt: Bridging policy learning and language modeling. *Advances in Neural Information Processing Systems*, 36:7216–7262, 2023.

Gower, J. C. and Dijksterhuis, G. B. *Procrustes Problems*, volume 30. OUP Oxford, 2004.

Hao, S., Gu, Y., Ma, H., Hong, J. J., Wang, Z., Wang, D. Z., and Hu, Z. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*, 2023.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hua, T., Yun, T., and Pavlick, E. mothello: When do cross-lingual representation alignment and cross-lingual transfer emerge in multilingual models? *arXiv preprint arXiv:2404.12444*, 2024.

Huh, M., Cheung, B., Wang, T., and Isola, P. Position: The Platonic representation hypothesis. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 21–27 Jul 2024. URL `https://proceedings.mlr.press/v235/huh24a.html`.

Joshi, A., Sharma, V., and Modi, A. CheckersGPT: Learning world models through language modeling. In *Annual Meeting of the Association for Computational Linguistics*, 2024. URL `https://api.semanticscholar.org/CorpusID:272779453`.

Karvonen, A. Emergent world models and latent variable estimation in chess-playing language models. *arXiv preprint arXiv:abs/2403.15498*, 2024. URL `https://api.semanticscholar.org/CorpusID:268681535`.

Karvonen, A., Wright, B., Rager, C., Angell, R., Brinkmann, J., Smith, L., Mayrink Verdun, C., Bau, D., and Marks, S. Measuring progress in dictionary learning for language model interpretability with board game models. *Advances in Neural Information Processing Systems*, 37:83091–83118, 2024.

Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*, 2017.

Li, J., Karamolegkou, A., Kementchedjhieva, Y., Abdou, M., and Søgaard, A. Structural similarities between language models and neural response measurements. In Sanborn, S., Shewmake, C., Azeglio, S., and Miolane, N. (eds.), *Proceedings of the 2nd NeurIPS Workshop on Symmetry and Geometry in Neural Representations*, volume 228 of *Proceedings of Machine Learning Research*, pp. 346–365. PMLR, 16 Dec 2024a. URL `https://proceedings.mlr.press/v228/li24a.html`.

Li, J., Kementchedjhieva, Y., Fierro, C., and Søgaard, A. Do vision and language models share concepts? a vector space alignment study. *Transactions of the Association for Computational Linguistics*, 12:1232–1249, 2024b.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. Emergent world representations: Exploring a sequence model trained on a synthetic task. *ICLR*, 2023.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., and Chang, K.-W. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

Merullo, J., Castricato, L., Eickhoff, C., and Pavlick, E. Linearly mapping from image to text space. In *The Eleventh*

International Conference on Learning Representations, 2023. URL https://openreview.net/forum?id=8tYRqb05pVn.

Mitchell, M. and Krakauer, D. C. The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120, 2023. doi: 10.1073/pnas.2215907120. URL https://www.pnas.org/doi/abs/10.1073/pnas.2215907120.

Mollo, D. C. and Millière, R. The vector grounding problem. *arXiv preprint arXiv:2304.01481*, 2023. URL https://arxiv.org/abs/2304.01481.

Nanda, N., Lee, A., and Wattenberg, M. Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*, 2023.

Ngo, J. and Kim, Y. What do language models hear? Probing for auditory representations in language models. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5435–5448, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.297. URL https://aclanthology.org/2024.acl-long.297/.

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., Botvinick, M., and Fedorenko, E. Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, 9(1):963, 2018.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.

Søgaard, A. Grounding the vector space of an octopus: Word meaning from raw text. *Minds and Machines*, 33 (1):33–54, 2023. doi: 10.1007/s11023-023-09622-4.

Spies, A. F., Edwards, W., Ivanitskiy, M. I., Skapars, A., Rauker, T., Inoue, K., Russo, A., and Shanahan, M. Transformers use causal world models in maze-solving tasks. *arXiv preprint arXiv:2412.11867*, 2024. URL https://api.semanticscholar.org/CorpusID:274776417.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Vafa, K., Chen, J., Rambachan, A., Kleinberg, J., and Mullainathan, S. Evaluating the world model implicit in a generative model. *Advances in Neural Information Processing Systems*, 37:26941–26975, 2024.

Yuan, Y. and Søgaard, A. Revisiting the Othello world model hypothesis. *arXiv preprint arXiv:2503.04421*, 2025. URL https://arxiv.org/abs/2503.04421.

Yun, T., Zeng, Z., Handa, K., Thapliyal, A. V., Pang, B., Pavlick, E., and Sun, C. Emergence of abstract state representations in embodied sequence modeling. *arXiv preprint arXiv:2311.02171*, 2023.

Zhuang, C., Fedorenko, E., and Andreas, J. Visual grounding helps learn word meanings in low-data regimes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1311–1329, 2024.

A. Potential Impacts

Our multi-modal Othello framework demonstrates how integrating visual and textual modalities can enhance structured reasoning in environments with strict rule-based dynamics. Beyond board games, this approach offers insights into multi-modal learning for tasks requiring spatial-temporal understanding, such as strategy modeling, robotics, and educational AI systems. By disentangling perceptual and symbolic reasoning, it also serves as a testbed for evaluating how models learn abstract rules from multi-modal input, potentially informing the design of more robust, interpretable, and generalizable multi-modal AI systems. Future work may generalize these insights to more complex domains and explore the role of other modalities, such as spatial or tactile input, in supporting the emergence of grounded representations.

B. Limitations

A notable limitation of this work is that we are not able to compare VISOTHELLO with autoregressive multi-modal large language models (MLLMs) due to fundamental differences in training paradigms. Autoregressive MLLMs treat images as part of a sequential token stream, effectively converting static visual-text inputs into video modeling tasks, which significantly increases computational complexity and alters the problem structure. In contrast, our model uses masked language modeling (MLM) to enable efficient bidirectional reasoning over static data, making direct comparison with autoregressive MLLMs infeasible without substantial task reformulation.

Moreover, we do not include comparisons with large-scale text-only language models, as these have been thoroughly investigated in prior work (Yuan & Søgaard, 2025). Given that pretraining on language alone does not necessarily enhance understanding of the structured reasoning inherent in Othello, scaling up to such models and benchmarking against them is not currently a priority. Instead, our use of lightweight language models offers a practical and efficient probe into how much language pretraining contributes to this domain.

C. Ethics Statement

We ensure that all datasets used in this work are publicly available and released under appropriate open-source licenses. No personal information about players or tournaments is included or revealed. Additionally, all corresponding images used in our experiments are synthetically generated, and do not depict real individuals or contain sensitive content.

D. Related Work

D.1. LLMs for Game Sequence Modeling

Using AI models to play games is not a new concept. Early models, such as AlphaGo, were designed to master gameplay by using predefined game rules and structured environments (Silver et al., 2016; 2017; Feng et al., 2023). Recently, modeling games with LLMs and examining their understanding of game dynamics has become a popular research direction in LLM cognitive probing. Li et al. (2023) train GPT-2 on synthetically generated Othello games, then use probing techniques to determine whether the model develops internal representations of the game state—effectively inferring a world model. Building on this work, Nanda et al. (2023) demonstrate that game-related knowledge is linearly encoded within the model. Following this line, research has expanded the scope of world knowledge acquisition in other scenarios with more advanced probing methods (Hao et al., 2023; Yun et al., 2023; Vafa et al., 2024). For instance, works train similar models with other game datasets, such as chess, maze and checkers, finding that the same encoding patterns hold in these more complex games (Karvonen, 2024; Spies et al., 2024; Joshi et al., 2024; Karvonen et al., 2024). More relevant to our work, Yuan & Søgaard (2025) extend the study beyond GPT-2, evaluating state-of-the-art LLMs (e.g., LLaMA-2 (Touvron et al., 2023), Qwen (Bai et al., 2023)) to assess their capacity for structured game knowledge representation. Hua et al. (2024) explore this phenomenon in multilingual settings, examining how language models encode and transfer game-related knowledge across different languages. Our work is the first to incorporate visual information in Othello game understanding, providing deeper insights into board state representations.

D.2. Multi-modal Alignment

A growing body of research explores cross-modal alignment as a lens to understand the extent to which language models can internalize and generalize knowledge from text-only inputs (Pereira et al., 2018; Caucheteux et al., 2022; Li et al., 2024a;

Split	Games	Images	Avg. per Game
Train	20,525	1,247,852	60.8
Validation	1,282	78,141	60.9
Test	3,850	233,975	60.8
Total	25,657	1,559,968	60.8

Table 2. Dataset statistics. The number of games and images per split. Each game comprises a sequence of steps, with one image per step.

Ngo & Kim, 2024). Notably, Merullo et al. (2023) demonstrate that visual representations can be effectively projected into the linguistic embedding space using simple linear transformations, revealing a surprising degree of structural compatibility between visual and textual modalities. Building on the theme, Li et al. (2024b) and Huh et al. (2024) argue that as model capacity increases, representations across modalities tend to converge toward a shared, modality-agnostic statistical structure of the world. Unlike prior work focused on aligning visual and linguistic representations of concrete objects, we extend this to abstract game mechanics, enabling deeper insight into how models understand structured environments from text alone.

E. Model Design Motivation

We employ a masked language modeling (MLM) objective, following VisualBERT, rather than an autoregressive GPT-style objective, primarily to reduce computational cost and model complexity. Autoregressive models require sequential token generation and unidirectional attention, which becomes inefficient when integrating visual features: inserting images into the token stream introduces a pseudo-temporal dimension, effectively treating the input as a **video** sequence. This not only increases memory and compute requirements but also complicates the training dynamics without offering clear benefits for static visual-text reasoning tasks such as Othello. In contrast, MLM supports bidirectional attention and parallel training, allowing the model to leverage both visual and textual context efficiently. By avoiding unnecessary autoregressive structure, we preserve modeling capacity while significantly reducing resource demands.

F. Baselines

Text-only models. We evaluate two text-only models with different architectures. (i) Othello-GPT, introduced by Li et al. (2023), is based on GPT-2 and trained autoregressively on Othello move sequences to predict the next move in a purely textual setting. (ii) BERT (Devlin et al., 2019) is trained using the same language learning objectives as VISOTHELLO, including both random token masking and future token masking. As BERT serves as the language backbone of our multi-modal model, it provides a strong baseline for isolating the contribution of visual information in learning Othello strategies.

Vision-only models. As a vision-only baseline, we train a ResNet-18 model (He et al., 2016) on board images. Unlike VISOTHELLO, which processes a sequence of board images and move tokens, the ResNet model is trained to predict the next move based solely on a single board image representing the current game state. It does not observe any move history or future states.

G. Model Training Details

All models are trained for up to 1000 epochs, with validation performed every 10 epochs. We apply early stopping with a patience of 5 validation steps, and retain the checkpoint with the highest validation accuracy for final evaluation. Training is conducted on a single NVIDIA A100-40GB GPU. BERT and VISOTHELLO are trained with a batch size of 128 and a learning rate of 1e-4, while ResNet is trained with a batch size of 512 using the same learning rate. Details of the dataset is shown in Table 2.

H. Rotation Demonstration

We demonstrate a board rotation example in Figure 4 .

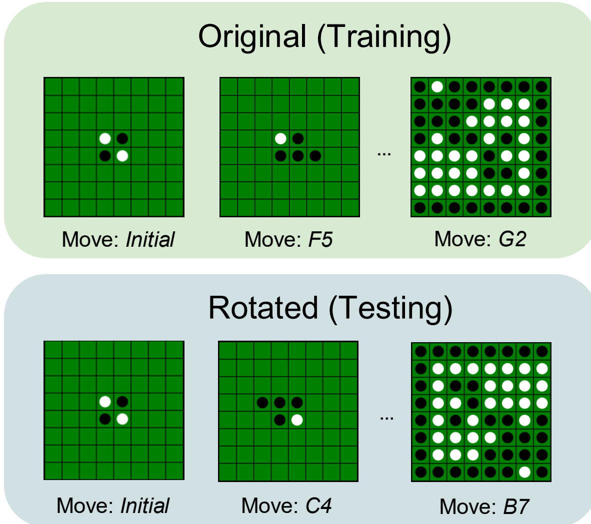


Figure 4. Illustration of Rotation 180°. A 180° rotation preserves game dynamics due to the board's inherent symmetry and the uniformity of move rules, making such transformations invariant under play.

Method	legal acc
VISOTHELLO	94.03
Pooling	92.43
Area	91.80
W/O FT ResNet	92.04
W/O FTM	62.03

Table 3. Ablation results for VISOTHELLO, using different image encoders (Pooling, Area, ResNet without fine-tuning) and without future token masking (W/O FTM). Results are based on the best validation checkpoints.

I. Ablation Study

We perform ablations to assess the impact of the modified components in VISOTHELLO relative to the original VisualBERT.

We assess whether fine-tuning a ResNet model for image feature extraction is necessary. Specifically, we compare it with simpler alternatives that do not involve task-specific training. The first approach (*Pooling*) performs a downsampling operation on the raw image of size $600 \times 600 \times 3$ using a spatial pooling mechanism, reducing it to a 1200-dimensional embedding. The second approach (*Area*) employs pixel area relation interpolation, where the image is partitioned into fixed-size patches, pixel values within each region are averaged, and the resulting resized image is flattened into a 1200-dimensional vector. We also evaluate a ResNet-18 encoder without any fine-tuning on Othello images.To test the role of future token masking, we ablate this component by training VISOTHELLO without it (denoted as *W/O FTM*) and compare performance against the full model.

According to Table 3, both the image encoder choice and future token masking significantly affect model performance. Replacing the fine-tuned ResNet with simpler alternatives such as Pooling or Area projections results in moderate drops in legal move accuracy (from 94.03% to 92.43% and 91.80%, respectively). Using ResNet without fine-tuning also leads to lower performance (92.04%), highlighting the benefit of domain-specific adaptation. Notably, removing future token masking causes a substantial degradation (62.03%), demonstrating its critical role in aligning the training objective with the causal structure of the game.

Source	Target	0	1k	3k	5k	10k	20k
BERT	ResNet	25.37	27.39	29.48	32.07	33.34	34.25
BERT	Othello-GPT	**86.01**	62.02	56.59	57.14	61.60	63.30
BERT	VISOTHELLO	83.92	61.76	54.83	53.96	55.59	57.94
Othello-GPT	ResNet	32.16	31.79	29.95	34.33	33.74	36.41
Othello-GPT	VISOTHELLO	83.09	**77.68**	**81.62**	**76.56**	**82.07**	**82.35**
VISOTHELLO	ResNet	11.62	26.03	29.04	33.83	32.49	38.99

Table 4. Supervised alignment similarity between target and source models. Highest in bold.

J. Feature Alignment

We perform representation alignment across models trained on Othello game sequences to assess whether models trained on different modalities (i.e., image and text) learn similar representations. Through this, we investigate whether modality-specific models encode analogous patterns that are fundamental to rule-following gameplay.

J.1. Alignment Method

We extract intermediate representations, denoted as H_i, from different models for alignment, using the same input sequence, and corresponding board images for multi-modal models. Specifically, we use the features extracted from final hidden layer of both encoder-only models (e.g., BERT, VISOTHELLO) and decoder-only models (e.g., Othello-GPT). Given the learned representations H_1 and H_2 of dimensions d_1 and d_2, respectively, extracted from models M_1 and M_2 based on the same game sequence input, we first apply PCA to project them into a shared-space of dimension $d = \min(d_1, d_2)$:

$$H_1' = P_d(H_1), H_2' = P_d(H_2), \tag{1}$$

where $H_1', H_2' \in \mathbb{R}^d$ are projected vectors.

Next, we align these representations into a common vector space using the MUSE package,[2] originally developed for mapping multilingual word embeddings into a shared space. The aim is to learn a linear mapping matrix W, for each projected representation H_1' and H_2'

$$W^* = \underset{W \in \mathcal{M}_i(\mathbb{R})}{\arg\min} \|H_i' W - H_j'\|, \tag{2}$$

where $i, j \in \{1, 2\}$ and $i \neq j$. This denotes learning the optimal linear mapping matrix W^* that aligns representation H_i' to H_j'.

J.2. Alignment Training

To obtain the optimal mapping matrix, we use both supervised and unsupervised training methods.

Supervised training. We treat representations from different models (e.g., Othello-GPT and VISOTHELLO) corresponding to the same game sequence as paired training data. For example, given the Othello move sequence input "*F5 F6 E6 F4 C3 D7*", the pairwise training input H_1' and H_2' correspond to the representations extracted from Othello-GPT and VISOTHELLO models, respectively, for this exact sequence and the associated images (when applicable). The mapping matrix W is learned and optimized with iterative Procrustes alignment (Gower & Dijksterhuis, 2004), which alternates between solving for the optimal orthogonal transformation and refining the mapping. This process minimizes the distance between the transformed source representations and the target representations, resulting in better alignment across the two vector spaces.

Unsupervised training. We also adopt the unsupervised training approach (Conneau et al., 2018; Lample et al., 2017) with the absence of paired data or predefined anchors to learn the alignment. Given a set of game features H' from both the source and target space, the process begins with adversarial training, where a discriminator is trained to distinguish whether the feature comes from the source or target representation space. Simultaneously, the mapping matrix W is optimized to make this distinction harder, effectively aligning the distributions. Once an initial mapping is obtained, we apply iterative Procrustes refinement, similar to the supervised setting, to improve the alignment. Alignment quality is evaluated and improved using the average cosine similarity between mapped source and target features on the test set.

[2] https://github.com/facebookresearch/MUSE

Source	**Target**	0	1k	3k	5k	10k	20k
BERT	ResNet	31.53	37.46	36.96	36.98	38.70	40.25
BERT	Othello-GPT	**90.38**	65.61	62.26	57.68	61.01	63.29
BERT	VISOTHELLO	90.52	67.52	62.23	58.97	63.60	63.77
Othello-GPT	ResNet	33.94	46.43	44.55	50.15	46.96	47.89
Othello-GPT	VISOTHELLO	87.20	**80.50**	**80.53**	**79.27**	**85.81**	**82.46**
VISOTHELLO	ResNet	23.04	43.44	44.39	45.38	52.64	57.79

Table 5. Unsupervised alignment similarity between target and source models. Highest in bold.

J.3. Alignment Training Setups

To construct the alignment training set, we randomly sample one subsequence from each complete game, resulting in 3,849 input sequences, each paired with the corresponding board state images. We then divide the data into training and testing sets with an 80%/20% split, resulting in 3,079 and 770 instances. We adopt cosine similarity to measure the alignment quality between representations from different models. After projecting the representations into a shared space, we compute the average pairwise cosine similarity between aligned feature vectors. A higher similarity score indicates better alignment, suggesting that the models, despite being trained on different modalities, capture similar underlying patterns. We train the alignment model using a single NVIDIA A100 GPU. All hyperparameters follow the default settings provided by the original MUSE implementation, with no additional tuning.

J.4. Mapping Result

Table 4 and 5 demonstrate the mapping results under supervised and unsupervised training. We find that the alignment similarity generally improves as the size of the training data increases. This trend suggests that with more data, the models learn richer and shared representations that are easier to align across modalities. Also, despite the difference in training strategy (i.e., autoregressive training and mask language modeling), Othello-GPT and BERT exhibit strong alignment, reflected in their high similarity scores. Surprisingly, Othello-GPT exhibits a strong alignment score with VISOTHELLO, indicating that despite differences in architecture and training modalities, the two models learn remarkably similar representations. This suggests that the underlying patterns essential for Othello gameplay are captured consistently across both language-based and multi-modal models. Such alignment highlights the potential for cross-modal knowledge transfer and opens avenues for further exploration of unified representations in complex tasks.